



Alliance Permanent Access to the
Records of Science in Europe Network

Quality Assurance of Research Data:

Perspectives of Scientist, Infrastructure
Providers, and Publishers

Heinz Pampel

Helmholtz Association, Helmholtz Open Access
Coordination Office

APARSEN Satellite Session to the Conference “Open
Access Tage 2012”, Vienna, 27 Sept. 2012



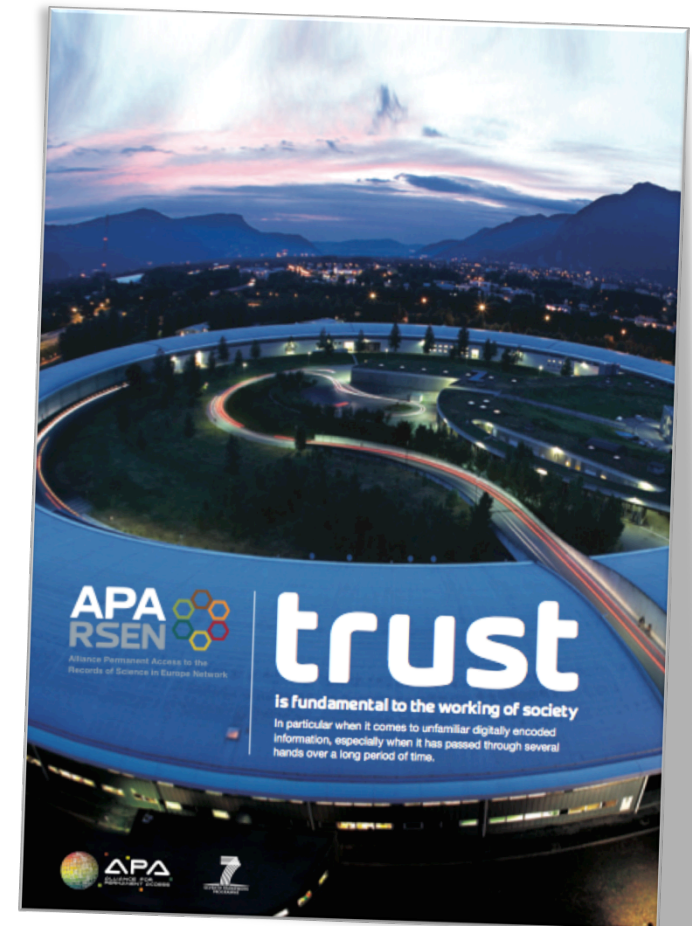
Co-ordinated by



Science & Technology
Facilities Council

Outline

- Quality: Definitions and Aspects
- The Scientist's Perspective
- The Data Repository's Perspective
- The Journal's Perspective
- Summary



EUROHORCs/ESF: ERA Vision, 2008, 2009

Visions

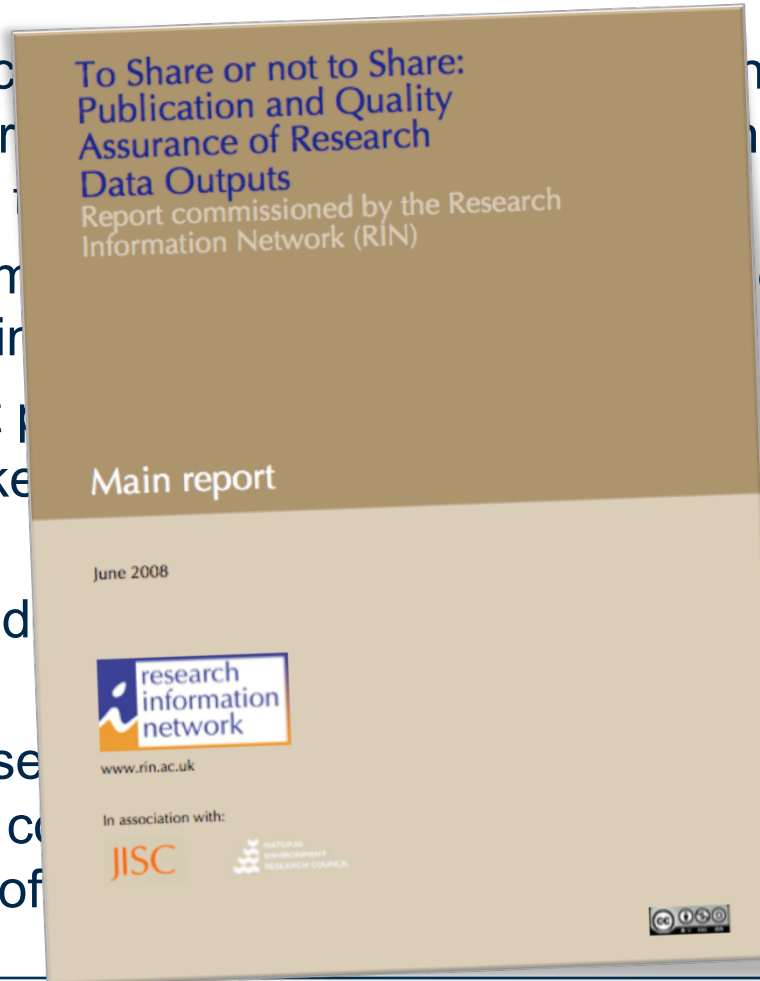
A globally competitive European Research Area (ERA) of excellence, to facilitate the advancement of science and help create a knowledge-based society in Europe, requires:

1. An effective European research policy, capitalising on cultural, geographic and scientific diversity;
2. A stimulating education system;
3. A single European labour market for researchers;
4. Adequate funding for top-quality, curiosity-driven research;
5. Transnational funding, benchmarking of quality and shared scientific priorities for strategic research and researcher-driven programmes;
6. Excellent research institutions;
7. World-class research infrastructures;
8. Open access to the output of publicly funded research and permanent access to primary quality-assured research data;
9. Effective and trusted bridges between science, society and the private sector;
10. Openness to the world.

- A globally competitive European Research Area (ERA) requires:
- “Open access to the output of publicly funded research and permanent access to primary quality-assured research data”

Research Information Network (RIN), 2008

- “The term “quality” is often used to mean the notion of being “fit for purpose”. With regard to research datasets we identified three key criteria:
 - first, the datasets must be of high quality, as defined by the data creators’ original intentions;
 - second, they must be of high quality, as defined by the work that has been undertaken to ensure that the data has been validated by other researchers;
 - third, they should identify the data as being suitable for re-use by others.
- Fulfilling the first and second criteria is a focus on scholarly method and content, while the third criterion focuses on the technical aspects of data management and sharing.



Research Information Network (RIN), 2008

- “The term “quality” is conventionally associated with the notion of being “fit for purpose”. With regard to creating, publishing and sharing datasets we identified three key purposes:
 - first, the datasets must meet the purpose of fulfilling the goals of the data creators’ original work;
 - second, they must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers;
 - third, they should ideally be discoverable, accessible and re-usable by others.
- Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated.“

Waijers & Van der Graaf, 2011

- Categorisation:
 - Quality assurance in the data creation process
 - Data management planning
 - Quality assessment of datasets



Thomas Hawk (CC-BY) on Flickr: <http://www.flickr.com/photos/thomashawk/3182986457>

UK: Science and Technology Committee, 2011



House of Commons
Science and Technology
Committee

Peer review in scientific publications

Eighth Report of Session 2010–12

*Volume I: Report, together with formal
minutes, oral and written evidence*

*Additional written evidence is contained in
Volume II, available on the Committee website
at www.parliament.uk/science*

*Ordered by the House of Commons
to be printed 18 July 2011*

4 Data management

178. In paragraphs 21-22 we discussed the need for reviewers to assess manuscripts to ensure that they are technically sound. One of the questions that arose in the course of this inquiry was, how far should reviewers be expected to go to assess technical soundness? In this chapter we discuss the feasibility of reviewing the underlying data behind research and how those data should be managed.

The need to review data

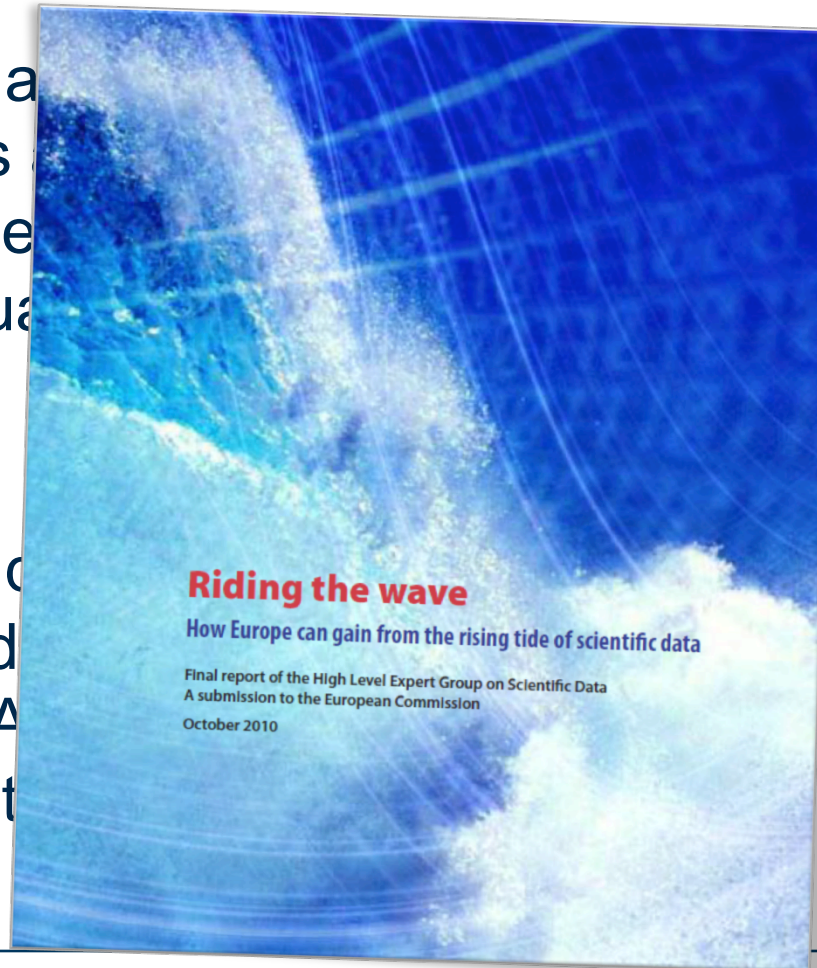
179. Sense About Science told us that:

The ultimate test of scientific data [...] comes through its independent replication by others; peer review is the system which allows publication of data so that it can be both criticised and replicated. It is a system which encourages people to ask questions about scientific data.³¹⁶

180. Replication does not usually take place during the peer-review process, although, “in exceptional circumstances, referees will undertake considerable work on their own initiative to replicate an aspect of a paper”.³¹⁷ Professor Sir Adrian Smith, Director General of Knowledge and Innovation in the Department for Business, Innovation and Skills (BIS).

HLEG on Scientific Data, 2010

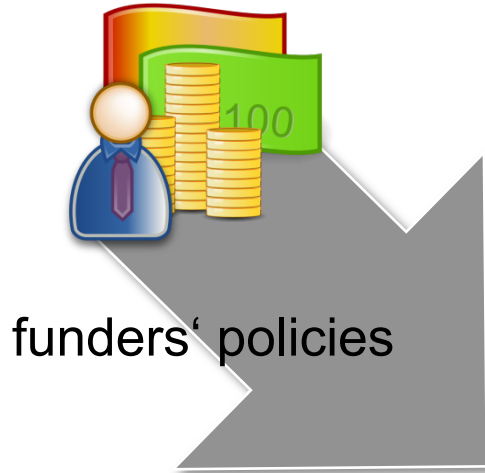
- „Researchers are able to find, access and use data with confidence in the discipline. They can be confident in the quality and data, and they can evaluate the data can be trusted.“
- “Producers of data want broad access, and prefer to deposit data in reliable evidence in reliable repositories. A discipline is guided by international standards that ensure data is trustworthy.”



HLEG on Scientific Data, 2010

- „Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.“
- “Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy.”

Stakeholder



scientists



data repositories



journals

[RRZE Icon Set](#) (CC: BY-SA)

The Scientist's Perspective

- Committee on Publication Ethics (COPE), 2008
 - “Reviewers should be asked to address ethical aspects of the submission such as: [...] Is there any indication that the data has been fabricated or inappropriately manipulated?”

- Research Information Network (RIN), 2008
 - “There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability.”

The Scientist's Perspective

- Mark Ware Consulting, 2008
 - **“A majority of reviewers (63%) and editors (68%) say that it is desirable in principle to review authors’ data.** Perhaps surprisingly, a majority of reviewers (albeit a small one, 51%) said that they would be prepared to review authors’ data themselves, compared to only 19% who disagreed. This was despite 40% of reviewers (and 45% of editors) saying that it was unrealistic to expect peer reviewers to review authors’ data. **Given that many reviewers also reported being overloaded, we wonder, however, whether they would still be as willing when it actually came to examine the data.”**

The Scientist's Perspective

- Sense about Science, 2009
 - “It is widely believed that peer review should act as a filter and select only the best manuscripts for publication. Many believe it should be able to detect fraud (79%) and plagiarised work (81%), but few have expectation that it is able to do this.
Comments from researchers suggest this is because reviewers are not in a position to detect fraud, this would require access to the raw data or re-doing the experiment.”
 - “[...] researchers point out that examining all raw data would mean peer review grinds to a halt.”

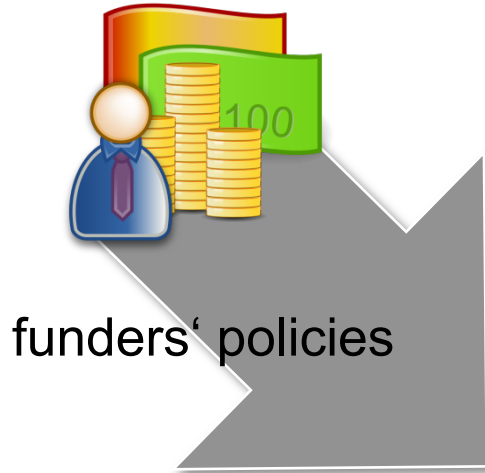
The Scientist's Perspective

- Waaijers & Van der Graaf, 2011
 - “Finally, it was suggested that, rather than setting up a separate quality assessment system for data, one could create a **citation system for datasets**, which would then form the basis for citation indices. The thinking behind this was that citation scores are a generally accepted yardstick for quality.”
 - “Scientists and scholars in all disciplines would welcome greater clarity regarding the re-use of their data, both through citations and through comments by re users. Setting up special **journals for data publications** is also popular in all disciplines.”
 - “The view regarding a mandatory section on data management in research proposals is also unanimous, but negative. The decisive factor here is a **fear of bureaucracy**.”

The Scientist's Perspective

- Key points
 - Scientists recognize that accessibility of data is a precondition for peer review of it.
 - In principle, reviewers and editors find it preferable for data to be peer reviewed but many reservations exist about its feasibility; „peer review may grind to a halt“.
 - Scientists fear that reviewing data in the course of the peer review process is not practical due to the amount of work and time involved.
 - Scientists have a positive attitude towards innovative publication strategies of research data and welcome greater clarity regarding the re-use of their data.
 - Scientists are sceptical about obligatory measures of data management, since they fear bureaucracy.

Stakeholder



scientists



data repositories



journals

[RRZE Icon Set](#) (CC: BY-SA)

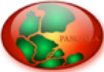
The Data Repository's Perspective

- e-IRG, 2009
 - “**Such digital data archives are the main advocates of quality assurance for research data.** Quality control by data archives is usually achieved by painstaking and labour-intensive checks on the data, carried out by data archive staff.”
- Research Information Network (RIN), 2011
 - “The curatorial role of the centre thus affects two important elements of data quality: first, **ensuring that individual datasets are academically „good”** (as much as it can) and second, ensuring that it **creates and preserves collections** which can be a useful starting point for new research.”

The Data Repository's Perspective

- Internal APARSEN survey:
 - The following measures of quality assurance were specified:
 - Business process documentation
 - Completeness / Consistency checks
 - Data curators technical review (methods, parameters, unit checks, consistency)
 - Data management and sharing training
 - File format validation
 - Metadata checks
 - Risk management
 - Storage integrity verification
 - Tools for annotating quality information

The Data Repository's Perspective



PANGAEA®
Data Publisher for Earth & Environmental Science

Not logged in (log in or sign up)

Always quote citation when using data!

[Show Map](#) [Google Earth](#) [RIS](#) [BisTeX](#)

Data Description

Citation: König-Langlo, G; Gernandt, H (2008): 426 ozonesonde profiles from Georg-Forster-Station. *Alfred Wegener Institute for Polar and Marine Research, Bremerhaven*, doi:10.1594/PANGAEA.547983,
Supplement to: König-Langlo, Gert; Gernandt, Hartwig (2009): Compilation of ozonesonde profiles from the Antarctic Georg-Forster-Station from 1985 to 1992. *Earth System Science Data*, 1(1), 1-5, doi:10.5194/essd-1-1-2009


Abstract: On 22 May 1985 the first balloon-borne ozonesonde was successfully launched by the staff of Georg-Forster-Station (70°46' S, 11°41' E). The following weekly ozone soundings mark the beginning of the continuous investigation of Germany to study the vertical ozone distribution in the southern hemisphere.
 In 1985 these ozone soundings have been the only record showing the change of vertical ozone distribution in the southern polar stratosphere in September and October. The regular ozone soundings from 1985 until 1992 are a valuable reference data set since the chemical ozone loss became a significant feature in the southern polar stratosphere.
 The balloon-borne soundings were performed at the upper air sounding facility of the neighbouring station Novolazarevskaya, just 2 km apart from Georg-Forster-Station. Till 1992, ozone soundings were taken without interruption. Afterwards, the ozone sounding program was moved to Neumayer-Station (70°39' S, 8°15' W) 750 km further west.

Project(s): [Meteorological Long-Term Observations @ AWI](#) (AWI_Meteo) ↗

Coverage: *Latitude:* -70.770000 * *Longitude:* 11.830000
Date/Time Start: 1985-05-22T05:19:00 * *Date/Time End:* 1992-01-29T01:19:00

Event(s): **GF_8771** ↗ * *Latitude:* -70.770000 * *Longitude:* 11.830000 * *Date/Time:* 1985-05-22T00:00:00 * *Location:* Antarctic ↗ * *Campaign:* Ozone_studies_1985-1992 ↗ * *Basis:* Georg Forster Station ↗ * *Device:* Radiosonde ↗
GF_8772 ↗ * *Latitude:* -70.770000 * *Longitude:* 11.830000 * *Date/Time:* 1985-05-24T00:00:00 * *Location:* Antarctic ↗ * *Campaign:* Ozone_studies_1985-1992 ↗ * *Basis:* Georg Forster Station ↗ * *Device:* Radiosonde ↗
GF_8773 ↗ * *Latitude:* -70.770000 * *Longitude:* 11.830000 * *Date/Time:* 1985-05-27T00:00:00 * *Location:* Antarctic ↗ * *Campaign:* Ozone_studies_1985-1992 ↗ * *Basis:* Georg Forster Station ↗ * *Device:* Radiosonde ↗

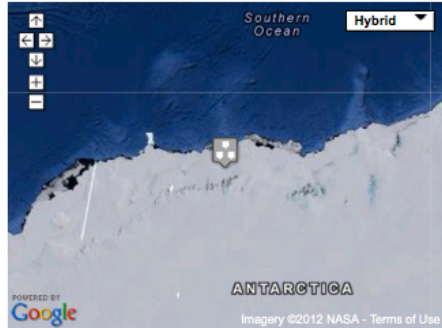
Comment: Attached to Russian radio sondes (type RKS-5) the ozone sondes (type OSE) were carried by balloons to heights up to 35 km. During the flight the measured ozone concentrations as well as the standard meteorological measurements were transmitted to the ground. All 426 soundings at the mean pressure levels and significant heights from these flights between 1985 and 1992 are archived in this dataset.
 The ozone measurements were achieved by a small electrically driven gas sampling pump which forces ambient air through a sensing solution of an electrochemical cell which generated an electrical current proportional to the mass flow rate of ozone. According to this principle (Brewer sonde), the sondes were developed and produced at the Akademiewerkstätten in East Berlin.

License:  Creative Commons Attribution 3.0 Unported

Size: 426 datasets

Download Data

Download **ZIP** file containing all datasets as tab-delimited text (use the following character encoding: ISO-8859-1: ISO Western (PANGAEA default))



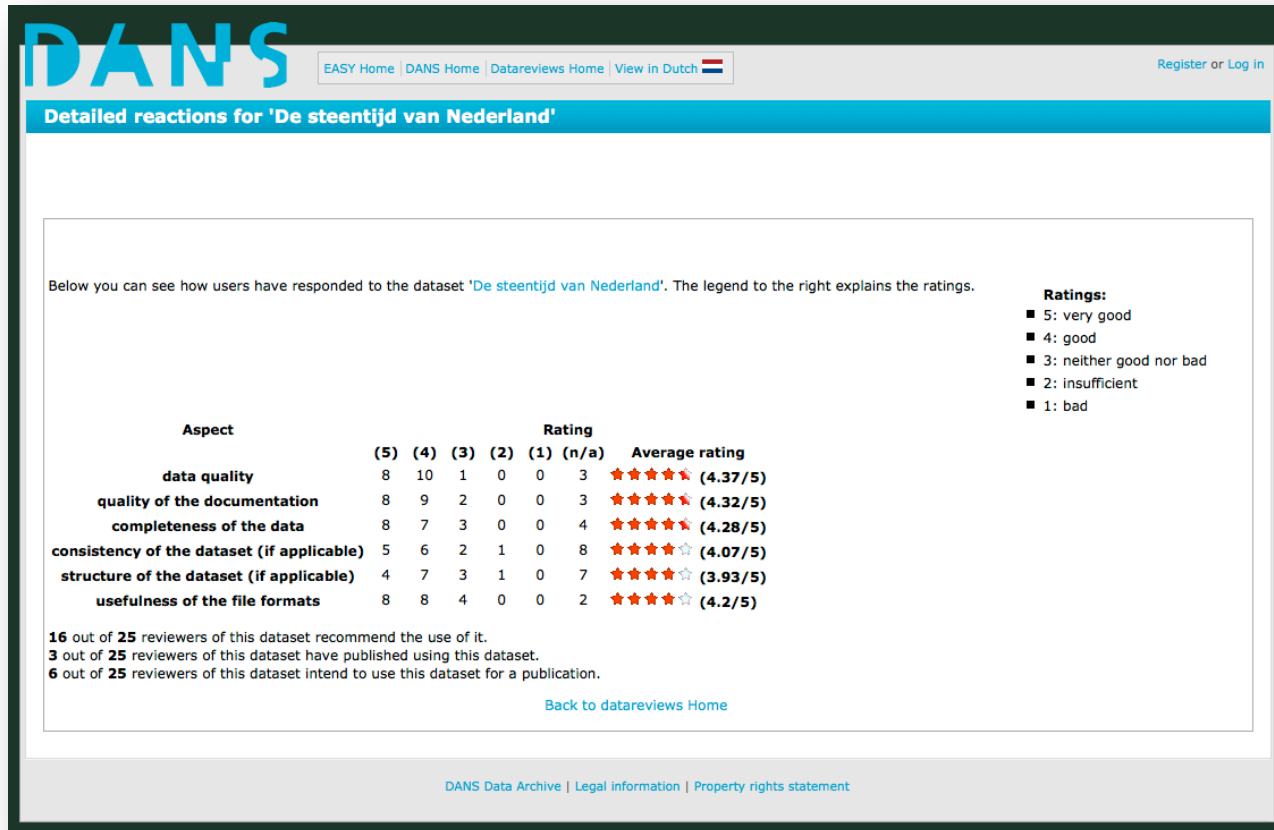
POWERED BY Google
Imagery ©2012 NASA - Terms of Use

The Data Repository's Perspective

- Example: PANGAEA
 - “The PANGAEA data **editorial** ensures the integrity and authenticity of your data. [...] The PANGAEA editors will check the **completeness and consistency** of metadata and data. Our editors are scientists from the earth and life sciences. We may **identify potential problems** with your data (e.g. outliers). Nevertheless, we will only take full responsibility for the **technical quality**. You will be responsible for the **scientific quality** of your data (e.g. the validity of used methods). After data have been archived you will receive a **DOI** name and you are requested to **proof-read** before the final version is published.”

The Data Repository's Perspective

- Example: DANS



The Data Repository's Perspective

• Example: DANS

Below you can see how users have responded to the dataset 'De steentijd van Nederland'. The legend to the right explains the ratings.

Ratings:

- 5: very good
- 4: good
- 3: neither good nor bad
- 2: insufficient
- 1: bad

| Aspect | Rating | | | | | | Average rating |
|--|--------|-----|-----|-----|-----|-------|----------------|
| | (5) | (4) | (3) | (2) | (1) | (n/a) | |
| data quality | 8 | 10 | 1 | 0 | 0 | 3 | ★★★★★ (4.37/5) |
| quality of the documentation | 8 | 9 | 2 | 0 | 0 | 3 | ★★★★★ (4.32/5) |
| completeness of the data | 8 | 7 | 3 | 0 | 0 | 4 | ★★★★★ (4.28/5) |
| consistency of the dataset (if applicable) | 5 | 6 | 2 | 1 | 0 | 8 | ★★★★★ (4.07/5) |
| structure of the dataset (if applicable) | 4 | 7 | 3 | 1 | 0 | 7 | ★★★★★ (3.93/5) |
| usefulness of the file formats | 8 | 8 | 4 | 0 | 0 | 2 | ★★★★★ (4.2/5) |

16 out of 25 reviewers of this dataset recommend the use of it.

3 out of 25 reviewers of this dataset have published using this dataset.

6 out of 25 reviewers of this dataset intend to use this dataset for a publication.

[Back to datareviews Home](#)

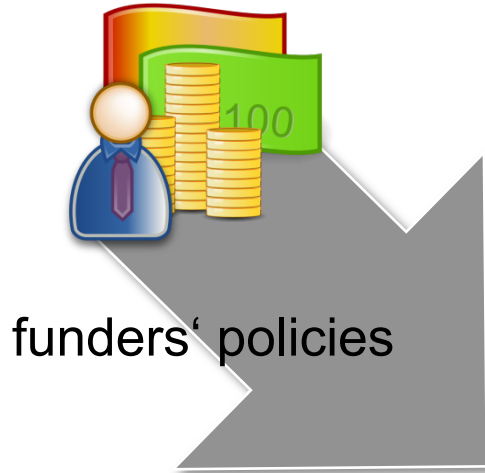
The Data Repository's Perspective

- Assessment and Certification
 - Data Seal of Approval
 - Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
 - DIN 31644 (Kriterien für vertrauenswürdige digitale Langzeitarchive)
 - DINI Certificate 2010 for Document and Publication Services
 - ISO-DIS 16363 (Audit and Certification of Trustworthy Digital Repositories)
 - ISO-DIS 16919 (Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Repositories)
 - Trustworthy Digital Repositories (RAC)
 - Trustworthy Repositories Audit & Certification (TRAC)
 - World Data System certification

The Data Repository's Perspective

- Key points:
 - Data repositories make a contribution to quality assurance of stored data.
 - Data management is assessed as an essential contribution to quality assurance of data. The selection process and subsequent verification of data (via persistent addressing) is seen as very important.
 - The measures contributed by repositories to quality assurance vary depending on the form, scope and discipline of data.
 - Certification and audit secure the quality of data repositories and affect the quality assurance of data.

Stakeholder



scientists



data repositories



journals

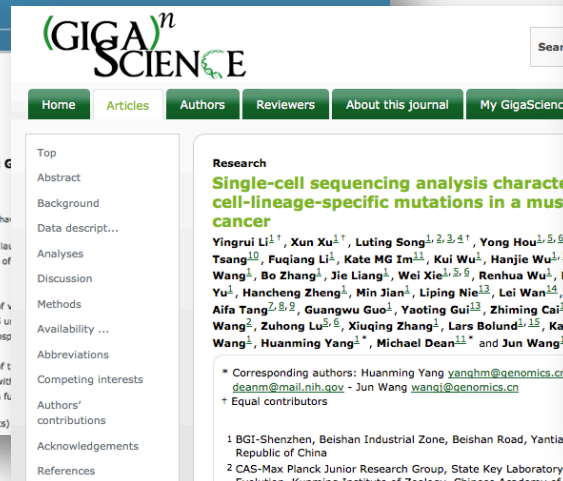
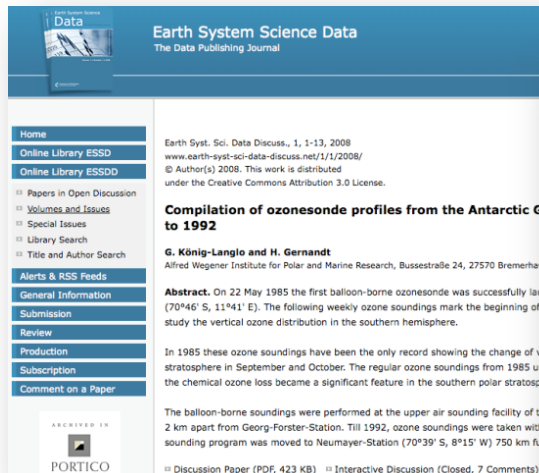
[RRZE Icon Set](#) (CC: BY-SA)

The Journal's Perspective

- Robert Campbell and Cliff Morgan of John Wiley & Sons:
 - “The real challenge is how to deal with the growth in research data that sits behind the journal article. Policies for data curation and sharing are emerging but there is no related peer review process or quality control.”
- Editorial policies:
 - Nature: “[...] condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without preconditions.”
 - PLoS: “PLoS is committed to ensuring the availability of data and materials that underpin any articles published in PLoS journals.”

The Journal's Perspective

- Data paper (Chavan & Penev, 2011)
 - “We define a data paper as a scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance to the standard academic practices.”



The Journal's Perspective

- APARSEN survey survey among editors and publishers
 - “The main challenges are to define the review criteria in a way that a non-paid reviewer is willing (not only is possible) to review the data and to reach the balance between the time to be spent to review data in depth on the one hand, but to keep the efforts for the review short on the other hand. **Reviewing data in depth is a great challenge. We have to find criteria and methods to allow reviewers to do a good review on data with moderate efforts and time.**” (response of a publisher)

The Journal's Perspective

- APARSEN survey survey among editors and publishers
 - “Data can only be reviewed properly when all underlying metadata, experiment conditions, etc. are fully shared with reviewers. This requires high standards on data sharing. **To share data and to review them is certainly beneficial to science, at the same time it puts additional strain on researchers.** This needs to be compensated with **incentives** (acknowledge the efforts for making data including appropriate metadata available; acknowledge the additional work in reviewing them).” (response of a publisher)

The Journal's Perspective

- APARSEN survey survey among editors and publishers:
 - Data papers:
 - “Where publication of a dataset is the primary purpose of a scholarly article, such as in the case of a data note, then it would be reasonable to infer a greater expectation of peer review of the related data.” (response of an editor)
 - “Its not entirely clear that reviewing a set of data without a paper is the same as reviewing a paper with claims/arguments built upon data.” (response of an editor)

The Journal's Perspective

- Key points:
 - Several journals require in their editorial policies the availability and accessibility of data, especially in the life sciences.
 - Peer review of underlying research data is not always included in the standard peer review process of journals.
 - In the peer review of publications, the main focus is on checking the claims and conclusions of the article. Peer review of underlying data plays a supportive role in this if and when useful to the reviewer.
 - In order to organize the reviewing of data effectively, clearly defined criteria are essential.
 - Publishers and editors have positive expectations of the development of data publications. They also expect that more in-depth peer review of data will take place for so-called data journals.

Summary

- **Scientists**

- Interdisciplinary exchange of methods of quality assurance of research data can help in disciplines which do not have fixed methods of establishing processes for quality assurance.
- Quality assurance of data is a time-consuming activity, which is not adequately recognized within scientific reputation systems. The development of incentive and reward systems can help to increase recognition for such work.

- **Data Repositories**

- To support scientists in quality assurance of data it is necessary to establish discipline-specific services of data management, which are in line with scientific requirements.
- The selection and verifiability of data in standardized form is attributed great importance within data management.
- Certification and audit secure the quality of data repositories and affect the quality assurance of data.

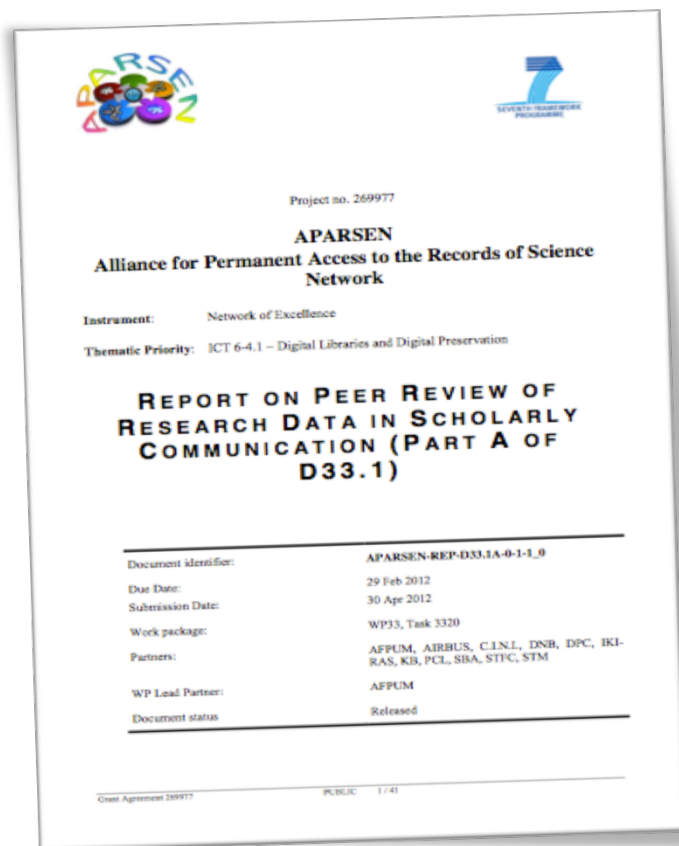
- **Journals**

- To organize reviewing of data effectively, standards and criteria of quality assurance have to be developed. Journals can make an important contribution here by formulating requirements of the quality of data in the editorial policies.
- Data publications provide a variety of opportunities of supporting the sharing of research data in a quality assured form.

References

- Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2
- Committee on Publication Ethics. (2008). A Short Guide to Ethical Editing for New Editors. Retrieved from http://www.publicationethics.org/files/short_guide_to_ethical_editing_for_new_editors.pdf
- Data Archiving and Networked Services. (2011). Data Reviews. Peer-reviewed research data. Retrieved from <http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5>
- Digital Curation Centre. (n.d.). Data Management Plans. Retrieved December 28, 2011, from <http://www.dcc.ac.uk/resources/data-management-plans>
- e-Infrastructure Reflection Group, & European Strategy Forum on Research Infrastructures. (2009). e- IRG Report on Data Management. Retrieved from http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf
- EUROHORCs & ESF. (2008). The EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for Actions to Help Build it. Retrieved from http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_ESF_ERA_RoadMap.pdf
- EUROHORCs & ESF. (2009). EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for Actions. Retrieved from http://www.era.gv.at/attach/EUROHORCs-ESF_Vision_and_RoadMap.pdf
- House of Commons. Science and Technology Committee. (2011). Peer review in scientific publications. Report, together with formal minutes, oral and written evidence. London. Retrieved from <http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsstech/856/856.pdf>
- Klump, J. (2011). Criteria for the Trustworthiness of Data Centres. *D-Lib Magazine*, 17(1/2). doi:10.1045/january2011-klump
- Mark Ware Consulting. (2008). Peer review in scholarly journals: Perspective of the scholarly community - an international study. Retrieved from <http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf>
- Morris, C. (Ed.). (1992). *Quality*. Academic Press Dictionary of Science and Technology. London: Academic Press.
- Pampel, H., Pfeiffenberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>
- PLoS ONE. (n.d.). PLoS ONE Editorial and Publishing Policies. Sharing of Materials, Methods, and Data. Retrieved December 28, 2011, from <http://www.plosone.org/static/policies.action#sharing>
- Pampel, H., & Bertelmann, R. (2011). „Data Policies— im Spannungsfeld zwischen Empfehlung und Verpflichtung. In S. Büttner, H.-C. Hobohm, & L. Müller (Eds.), *Handbuch Forschungsdatenmanagement* (pp. 49-61). Bad Honnef: Bock + Herchen. Retrieved from <http://opus.kobv.de/fhpotsdam/volltexte/2011/228/>
- Pfeiffenberger, H., & Carlson, D. (2011). —Earth System Science Data (ESSD) — A Peer Reviewed Journal for Publication of Data. *D-Lib Magazine*, 17(1/2). doi:10.1045/january2011-pfeiffenberger
- Research Information Network. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report. Retrieved from <http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf>
- Research Information Network. (2011). Data centres: their use, value and impact. Retrieved from [http://www.jisc.ac.uk/news/stories/2011/09/~media/Data Centres-Updated.ashx](http://www.jisc.ac.uk/news/stories/2011/09/~media/Data%20Centres-Updated.ashx)
- Ware, M. (2011). Peer Review: Recent Experience and Future Directions. *New Review of Information Networking*, 16(1), 23-53. doi:10.1080/13614576.2011.566812

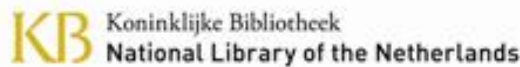
APARSEN Report



- Pampel, H., Pfeiffenberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>



Network of Excellence



CC-BY

- These slides are licensed under the Creative Commons „Attribution 2.0 Germany (CC BY 2.0)“ License. To view a copy of this license, visit: <http://creativecommons.org/licenses/by/2.0/>

